

Universidade Federal do Paraná
Setor de Ciências Exatas
Departamento de Estatística
Programa de Especialização em *Data Science* e *Big Data*

Paulo Candeia de Freitas Chixaro Neto

**Análise de Dados de Telemetria de Veículos
Focada na Redução de Manutenções
Corretivas Usando GAMLSS**

**Curitiba
2020**

Paulo Candeia de Freitas Chixaro Neto

Análise de Dados de Telemetria de Veículos Focada na Redução de Manutenções Corretivas Usando GAMLSS

Monografia apresentada ao Programa de Especialização em Data Science e Big Data da Universidade Federal do Paraná como requisito parcial para a obtenção do grau de especialista.

Orientador: Prof. Cesar Augusto Taconeli

Curitiba
2020

Análise de Dados de Telemetria de Veículos Focada na Redução de Manutenções Corretivas Usando GAMLSS

Paulo C. F. Chixaro Neto¹
Prof. Cesar Augusto Taconeli²

Resumo

Baseado em um sistema de telemetria veicular, que reflete o comportamento de um motorista, e também informações como hodômetro e idade do veículo, foi proposta uma regressão do tipo Modelo Aditivo Generalizado para Posição, Escala e Forma (GAMLSS) para analisar a frequência de manutenções corretivas. Utilizando-se a ferramenta R com o auxílio da biblioteca GAMLSS [1] foram ajustados diversos modelos baseados em distribuições para dados discretos e comparados através do Critério de Informação Akaike (AIC). O modelo com menor AIC usou a distribuição Binomial Negativa Inflacionada de Zeros (ZINBI) e alguns suavizadores do tipo *cubic splines*. O resultado final mostrou uma relação direta entre a frequência de manutenções corretivas e o valor do hodômetro e idade, o que era esperado. Essa mesma relação foi observada com a quantidade de registros de veículo ligado sem movimento.

Palavras-chave: telemetria, telemática, GAMLSS, regressão, manutenção corretiva.

Abstract

Based on telematics system, representing driver behaviour, also odometer and vehicle age, a Generalized Additive Model for Location, Scale and Shape was proposed having the number of repair services as response. Using software R and GAMLSS package some models with distinct distributions for distinct data were fitted and ranked by their Akaike Information Criterion values. The best model with the smallest value had a Zero Inflated Negative Binomial (ZINBI) distribution as family parameter e cubic splines smoothers. The final model summary showed as expected a direct relation between the response, repair services frequency, and odometer and age. This same relationship is also observed for quantity of idling events.

Keywords: telemetry, telematics, GAMLSS, regression, repair service.

1 Introdução

Um sistema de telemetria ou simplesmente telemática - combinação do prefixo grego 'tele' que significa 'a distância' e 'mática' da palavra 'informática', o processamento de informação - permite o recebimento de dados de um ou mais veículos por uma central, visando o monitoramento de ações do condutor e das informações sobre o veículo. Pode ser usado com uma série de propósitos comerciais e não comerciais, porém quando usado no gerenciamento de frotas pode trazer grandes vantagens para o profissional responsável por esses veículos, o gerente de frotas.[2]

Um veículo com sistema de telemetria pode enviar os dados colhidos por seus sensores para uma central em tempo real, enquanto está em trânsito (sistema ativo), ou descarregá-los para um computador toda vez que voltar para a base (sistema passivo). A comunicação ativa pode ser feita através de rede de telefonia móvel, via satélite ou ainda uma mistura dos dois, o que pode impactar no preço e disponibilidade do serviço. [3]

Ao ligar um veículo é registrado, no dispositivo de telemetria instalado, um código identificador de viagem (ID da viagem), assim todas as informações geradas enquanto o carro estiver ligado estarão relacionadas a esse código, por exemplo data e hora, latitude e longitude iniciais e finais da viagem, distância percorrida, velocidade média, velocidade máxima, nível de óleo do motor, temperatura da água e todos os outros dados colhidos por sensores instalados.

Um veículo ao longo de sua vida útil é submetido a manutenções preventivas (as revisões periódicas que ocorrem normalmente em intervalos de 5 a 10 mil quilômetros ou a cada 6 meses ou 1 ano) e corretivas (reparos não programados, como troca prematura de peças, serviços de funilaria e pintura). Esse segundo tipo pode ocorrer devido a terceiros ou à própria condução descuidada ou mesmo agressiva do condutor. Tal comportamento pode ser monitorado, via sistema de telemetria, através de dados representativos de superações do limite de velocidade das vias, manobras bruscas de aceleração, frenagem e curva, períodos prolongados em rotação excessiva do motor ou em ociosidade (motor ligado sem movimentação do veículo). Também pode ser acrescentado a esse sistema o número de multas obtidas no período observado categorizadas conforme a gravidade:

¹ Aluno do programa de Especialização em Data Science & Big Data, paulochixaro@gmail.com.

² Professor do Departamento de Estatística - DEST/UFPR.

leve, média, grave e gravíssima.

Manutenções corretivas representam uma despesa extra no planejamento de uma empresa que possui uma frota de veículos. Pensando nisso foi proposto desenvolver um modelo de regressão que pudesse se ajustar aos dados de telemetria e que como resposta tivesse a frequência de manutenções corretivas de determinado veículo. Tal proposta poderia ser oferecida como um produto para o gerente de frotas da empresa, visando auxiliá-lo na redução de custos da sua área. O tipo de regressão escolhida foi o Modelo Aditivo Generalizado para Posição, Escala e Forma (GAMLSS)[1] por ser muito mais flexível que um Modelo Linear Generalizado (GLM), contando com uma gama maior de distribuições, parâmetros para ajuste de inflação de zeros e ainda termos aditivos como suavizadores.

Os veículos analisados nesse artigo pertencem a uma empresa de aluguel de frotas focada em pessoas jurídicas. Para todos os contratos são fornecidos veículos novos, comprados especificamente para esse fim e ao término do período contratado os carros devolvidos são vendidos. A empresa locadora, bem como os dados utilizados serão mantidos anônimos ao longo do trabalho.

Na seção 2 são apresentados detalhadamente os dados trabalhados e as formas de medição de cada um deles. Para alguns dados foram aplicadas transformações, também explicadas nessa parte do trabalho. Subsequentemente, na seção 3, é mostrado um comparativo de diversos modelos ajustados com uso de diferentes distribuições e seus respectivos resultados. Mais adiante, na seção 4, encontra-se o ajuste fino do modelo utilizando a seleção de covariáveis em cada parâmetro e suavizadores. Na sequência estão ainda as análises dos gráficos de resíduos. Na última seção 5 estão algumas considerações finais sobre o resultado do trabalho e possíveis pontos de melhoria.

2 Descrição e Análise dos Dados

Os dados de telemetria utilizados foram obtidos através de 4 fornecedores de forma automatizada, desde o recebimento até a disponibilização via banco SQL. Informações que não necessitam de transmissão pelo sistema de telemetria como marca e modelo do veículo, quantidade e gravidade de multas, e registro de manutenções preventivas e corretivas (variável resposta e histograma mostrado na Figura 1), são registradas no sistema da empresa. O histórico delas, utilizado nesse trabalho, foi inserido manualmente ao final do fluxo de tratamento dos dados, que não será abordado neste artigo.

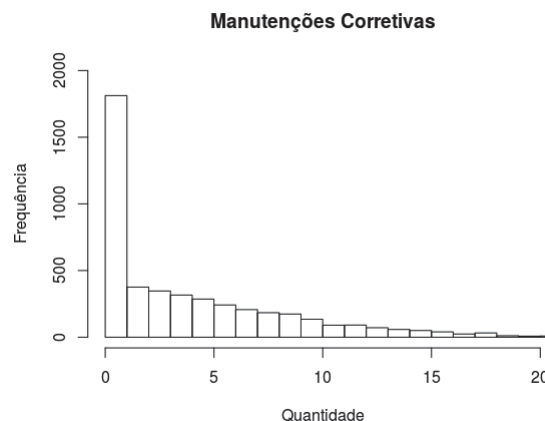


Figura 1: Histograma da quantidade de manutenções corretivas, variável resposta utilizada nos modelos.

As informações vindas dos fornecedores foram: hodômetro (quilometragem registrada no painel), número de eventos no período analisado e data da primeira e última informação enviada (utilizadas para determinar as idades dos veículos, pois todos eles tiveram telemetria instalada desde o primeiro uso). Os eventos foram divididos em: aceleração, curva e frenagem bruscas, excesso de velocidade e de rotações por minuto e motor ligado sem movimento.

Um evento brusco é registrado toda vez que o acelerômetro do veículo ultrapassa o valor de força G parametrizado lateral e longitudinalmente. O excesso de velocidade e de rotações por minuto, assim como o motor ligado sem movimento, são contados a cada vez que o veículo fica mais que determinado tempo acima dos valores estipulados. Cada fornecedor pode configurar os respectivos sensores como preferir ou de acordo com o desejado pelo cliente, mas em linhas gerais os valores são padronizados.

Um outro dado utilizado vindo dos fornecedores foi o número de viagens realizadas e a distância total percorrida em cada uma, separadas nos tipos "estrada" e "urbana".

As categorias dos carros foram inseridas de forma manual através de consulta de marca e modelo em site especializado [4] e separação nos grupos principais, de maior frequência na base: *Hatch* Compacto, *Pickup* Compacta, *Sedan* Compacto, *Furgão* Compacto, *Pickup* Média e Outras. Devido ao grande desbalanceamento nos dados, foi escolhida apenas a categoria *Hatch* Compacto, responsável por 71% dos dados coletados. Além disso não se pode esperar que as peças de um hatch compacto se comportem igualmente às de uma pickup média. Por serem veículos produzidos para usos diferentes possuem durabilidades diferentes de componentes, outro fator responsável pela decisão de seguir com apenas uma categoria.

Como dito anteriormente nessa seção pode haver diferença na configuração de alguns sensores entre fornecedores de telemetria distintos, portanto foi selecionado apenas o número 1, já que pela Tabela 1 é o fornecedor

com maior número de veículos na base.

Fornecedor	Frequência
1	2263
3	1258
4	797
2	317

Tabela 1: Frequência de cada fornecedor na base de dados.

Não existe registro informando o condutor de cada viagem, por isso cada caso foi tratado olhando-se apenas o veículo como identificador único da observação. Embora saibamos que o comportamento dos condutores pode variar muito, o trabalho ficou limitado em extrair os dados para cada veículo.

O número de manutenções preventivas foi dicotomizado de maneira a considerar se o veículo está com as manutenções em dia ou não. Essa covariável foi obtida comparando-se o número de manutenções preventivas com o hodômetro. Devido ao fato de existirem na base de dados diversas marcas e modelos de veículos, foi padronizada a necessidade de uma manutenção preventiva a cada dez mil quilômetros. Por exemplo um veículo que já rodou vinte e cinco mil quilômetros deveria ter realizado duas manutenções preventivas em sua vida, logo se realizou duas ou mais, a covariável "prev_ok" assume valor *TRUE*, caso contrário *FALSE*. De acordo com a Figura 2, relação entre a resposta e a nova covariável dicotomizada, o grupo dos veículos com as manutenções preventivas em dia possui uma mediana inferior ao outro e uma menor dispersão no número de corretivas, sendo um bom indício de ter sido uma covariável útil para os ajustes dos modelos.

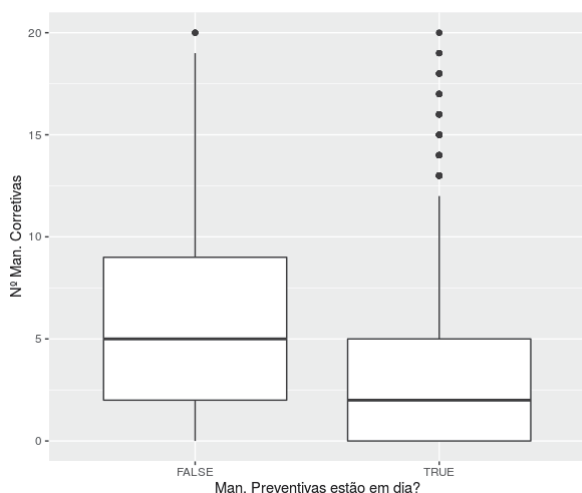


Figura 2: Boxplot do número de manutenções corretivas pelas manutenções preventivas em dia.

Outra covariável construída foi a proporção de multas (graves e gravíssimas) sobre o total de multas recebidas, com o objetivo de passar para os modelos uma medida que pudesse comparar um cenário de muitas multas

leves (razão próxima de zero) contra um de poucas multas porém sendo a maioria grave ou gravíssima (razão próxima de um).

Como o hodômetro do veículo é composto pela soma das distâncias de todas as viagens realizadas, não seria sensato carregar os dados de distância percorrida das viagens para o modelo, devido a problemas de convergência e multicolinearidade. Outro dado descartado foi a divisão das viagens em "estrada" e "urbana" devido ao severo desbalanceamento entre as duas, sendo 97% dos dados do tipo "estrada". Para as covariáveis de eventos e hodômetro ainda foi aplicada a transformação *log* devido às escalas serem muito discrepantes do restante e apresentarem forte assimetria. Para ajudar o leitor a acompanhar as análises do modelo da seção 3, na Tabela 2 está a descrição das covariáveis e da resposta utilizadas.

Campo	Descrição
corretiva	Número de manutenções corretivas
idade	Idade do veículo em dias
log_ace	Eventos de aceleração brusca
log_rpm	Eventos de excesso de RPM
log_vel	Eventos de excesso de velocidade
log_fre	Eventos de frenagem brusca
log_mov	Eventos de motor ligado s/ movimento
log_hodom	Hodômetro
qtdmultas	Quantidade de multas
taxa_grave	(Multas graves+gravíssimas)/qtdmultas
prev_ok	Manutenções preventivas em dia

Tabela 2: Descrição das covariáveis utilizadas para ajuste do modelo.

3 Escolha da Distribuição

Com as covariáveis devidamente selecionadas e as linhas filtradas em fornecedor e categoria, foram ajustados modelos da família GAMLSS utilizando algumas distribuições para dados de contagens, sendo encontradas e explicadas detalhadamente no documento de distribuições para uso no pacote GAMLSS [5].

Um simples comparativo de cálculo de média e variância da quantidade de manutenções corretivas mostra uma discrepância entre a ordem de grandeza das duas medidas, sendo a média igual a 4,97 enquanto a variância é 26,72, mais de 5 vezes maior, portanto um indicador de que pode existir uma superdispersão nos dados da base.

O primeiro teste foi a comparação entre o uso das distribuições Poisson (PO) e Binomial Negativa (NBI), utilizando todas as covariáveis para modelagem da média μ . Como a distribuição da Binomial Negativa possui um parâmetro a mais σ , para ajuste da dispersão, esse foi deixado constante numa primeira etapa e depois foi configurado sua dependência em todas as covariáveis.

Na Figura 1 é possível observar que a frequência de zeros, representada na barra mais a esquerda é mais de três vezes maior que a segunda barra, portanto apontou

uma necessidade de uma distribuição que contemplasse isso, logo uma que tivesse um parâmetro para a inflação de zeros. Foi testada uma versão alternativa da família Binomial Negativa, a inflacionada de zeros (ZINBI), a fim de melhorar o ajuste. Para realizar o comparativo com as duas anteriores o parâmetro de dispersão σ foi deixado igual a da NBI, com dependência em todas as covariáveis, e o parâmetro que controla a inflação de zeros ν para a ZINBI, foi ajustado também dependendo de todas as covariáveis.

O método escolhido de comparação dos resultados foi o critério de informação de Akaike, AIC, calculado por $-2l(\hat{\theta}) + 2df$, onde $l(\hat{\theta})$ é a máxima log-verossimilhança do modelo e df é o número de graus de liberdade.

Pelos resultados apresentados na Tabela 3 vemos que a distribuição Poisson teve AIC maior que a Binomial Negativa e um dos fatores que levaram a isso foi essa superdispersão, que não é contemplada pela Poisson. Na distribuição NBI a relação entre variância e média é quadrática, permitindo acomodar a superdispersão. Também é possível ver que o parâmetro de inflação de zeros contribuiu para um AIC ainda menor, comparado com o modelo correspondente sem o parâmetro.

Família	AIC
ZINBI	4993,711
NBI c/ par	5017,874
NBI s/ par	5092,341
PO	5244,376

Tabela 3: AIC para cada modelo ajustado usando as famílias Poisson, Binomial Negativa e Binomial Negativa Inflacionada de Zeros).

O modelo com menor AIC da Tabela 3, distribuição Binomial Negativa Inflacionada de Zeros (ZINBI), se mostrou a melhor escolha para o objetivo desse trabalho, então com base nele foram feitos os ajustes mais finos em seus parâmetros para tentar melhorar o ajuste.

4 Ajuste do Modelo

O próximo passo foi aplicar o método *stepwise* para definição das covariáveis utilizadas na modelagem dos parâmetros do modelo. Foi utilizada a função *stepAIC* pertencente ao pacote *GAMLSS* para ajudar na escolha da melhor configuração. Ela é baseada na função *stepAIC*(), explicada no capítulo 7.2 em [1], e utiliza uma estratégia que consiste em aplicar o método *forward stepwise* para os parâmetros μ , σ e ν , e depois o método *backward stepwise* em σ e μ respectivamente.

No argumento *method* [1], responsável pelo algoritmo de convergência do modelo, foi utilizado um número maior de iterações, 100, pois com o valor *default* não houve convergência. O modelo resultante da função *stepAIC*() foi nomeado "ZINBI_step" e as análises dos três parâmetros e seus resumos se encontram a seguir.

Para o parâmetro da média μ , Figura 3, as covariáveis escolhidas pelo método *stepwise* são estatisticamente significativas, por apresentarem valor p menor ou igual a 0,05. Também é possível ver a relação direta entre as covariáveis e a frequência de manutenções corretivas, por exemplo quanto maior a idade ou hodômetro de um veículo, mais manutenções corretivas ele tende a ter.

```
Mu link function: log
Mu Coefficients:
      Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.031e+01  4.440e-01 -23.228 < 2e-16 ***
log_hodom    9.724e-01  4.146e-02  23.454 < 2e-16 ***
idade        7.285e-04  7.560e-05  9.637 < 2e-16 ***
log_mov      1.090e-01  1.334e-02  8.174 7.12e-16 ***
prev_okTRUE  2.166e-01  3.741e-02  5.789 8.89e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figura 3: Resumo do ajuste do modelo para os termos do parâmetro de média μ .

No parâmetro σ , Figura 4, é possível dizer que a covariável "prev_ok" se mostrou menos relevante estatisticamente, valor p 0,0685, e posteriormente o modelo foi ajustado sem ela.

```
Sigma link function: log
Sigma Coefficients:
      Estimate Std. Error t value Pr(>|t|)
(Intercept)  13.3864    1.6266   8.230 4.6e-16 ***
log_hodom    -1.3927    0.1446  -9.634 < 2e-16 ***
prev_okTRUE  -0.4813    0.2640  -1.823 0.0685 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figura 4: Resumo do ajuste do modelo para os termos do parâmetro de dispersão σ .

Finalmente para o parâmetro que controla a inflação de zeros, ν , Figura 5 as covariáveis apresentaram valores p acima de 0,05, sendo "idade" o mais significativo deles com 0,095.

```
Nu link function: logit
Nu Coefficients:
      Estimate Std. Error t value Pr(>|t|)
(Intercept)  11.77746    6.83587   1.723 0.0852 .
idade       -0.03534    0.02115  -1.671 0.0950 .
qtdmultas    0.99852    0.69913   1.428 0.1535
prev_okTRUE  -9.62563    6.64121  -1.449 0.1475
log_vel     -1.83175    1.18712  -1.543 0.1231
taxa_grave  -255.73813   196.52799  -1.301 0.1934
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figura 5: Resumo do ajuste do modelo para os termos do parâmetro de inflação de zeros ν .

Foi avaliada a inclusão de um termo aditivo suavizador *cs()*, *cubic spline* [7], em todas as covariáveis do tipo numérico, a fim de encontrar alguma relação não linear nos preditores que pudesse ser melhorada com o uso do suavizador e resultar em um AIC menor. O melhor resultado foi obtido em "log_hodom" e "log_mov" no parâmetro μ , "log_hodom" no parâmetro σ e "taxa_grave" no parâmetro ν .

O campo "prev_ok" foi removido do preditor da dispersão por continuar com significância baixa após a aplicação dos suavizadores. Por outro lado as covariáveis da inflação de zeros tiveram uma melhora na significância, portanto foram mantidas nesse parâmetro do modelo.

Para análise do resultado do uso dos suavizadores foram gerados gráficos do tipo *term-plots*, ver [1], para visualizar se realmente existia uma relação não linear entre os preditores e a resposta. É possível observar que nos parâmetros de média (Figura 6), dispersão (Figura 7) e inflação de zeros (Figura 8) as covariáveis que receberam suavizadores apresentaram curvas acentuadas indicando de fato essa não linearidade com a variável resposta.

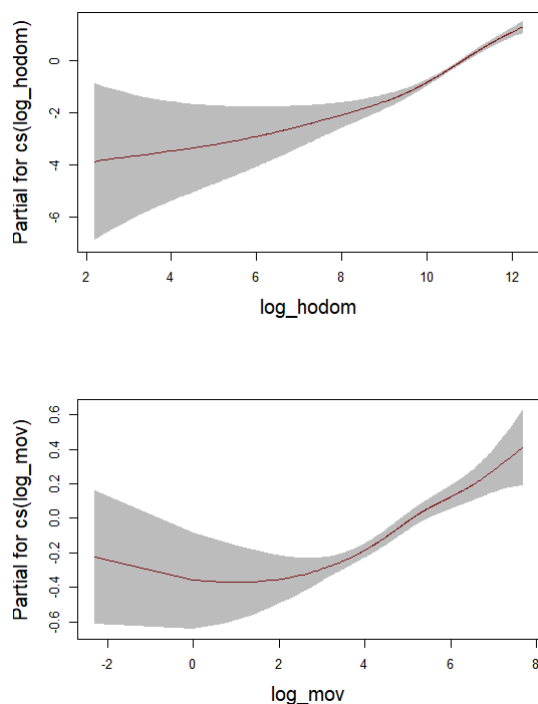


Figura 6: *Term-plot* do parâmetro de média μ .

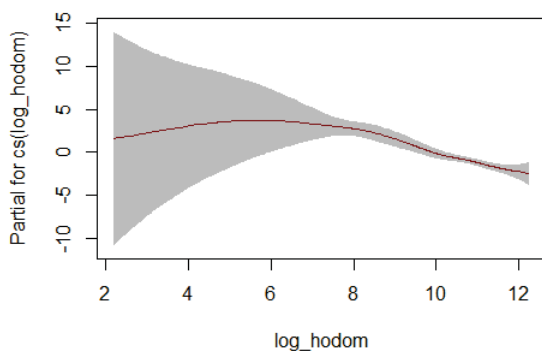


Figura 7: *Term-plot* do parâmetro de dispersão σ .

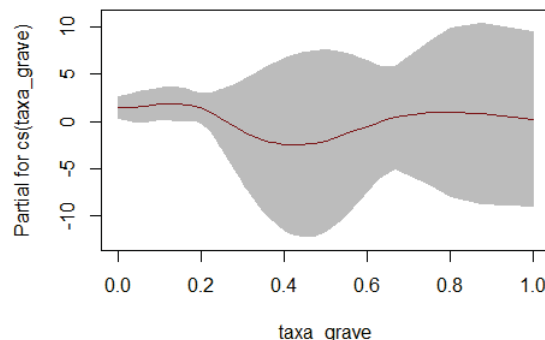


Figura 8: *Term-plot* do parâmetro de inflação de zeros ν .

O comparativo dos valores AIC do modelo final, após o uso do suavizador (mod_final), com o modelo após o método *stepwise* pode ser visto na tabela 4.

Modelo	AIC
mod_final	4959,285
ZINBI_step	4973,865

Tabela 4: AIC do modelo pós método *stepwise* e do modelo final, sem as covariáveis menos significativas e com termos suavizadores.

As funções *plot()* e *wp()*, *worm plot*, do pacote GAMLSS [1] geram alguns gráficos de resíduos interessantes para análise do ajuste do modelo. Na figura 9 é possível ver que no primeiro e segundo gráficos os resíduos quantílicos estão distribuídos aleatoriamente com variância aproximadamente constante ao longo do eixo y. A densidade dos resíduos observada no terceiro gráfico se aproxima muito de uma distribuição normal centrada em zero, sendo mais um indício de bom ajuste do modelo aos dados. Por fim, o *Q-Q Plot* possui a maioria dos pontos sobre a reta de identidade, com uma pequena fuga nas extremidades que não representa uma preocupação para o ajuste final.

O outro gráfico citado, derivado da função *wp()* Figura 10, mostra os pontos entre as bandas de intervalo de confiança de 95%, com alguns deles ultrapassando levemente as bandas, mostrando novamente um bom ajuste do modelo aos dados.

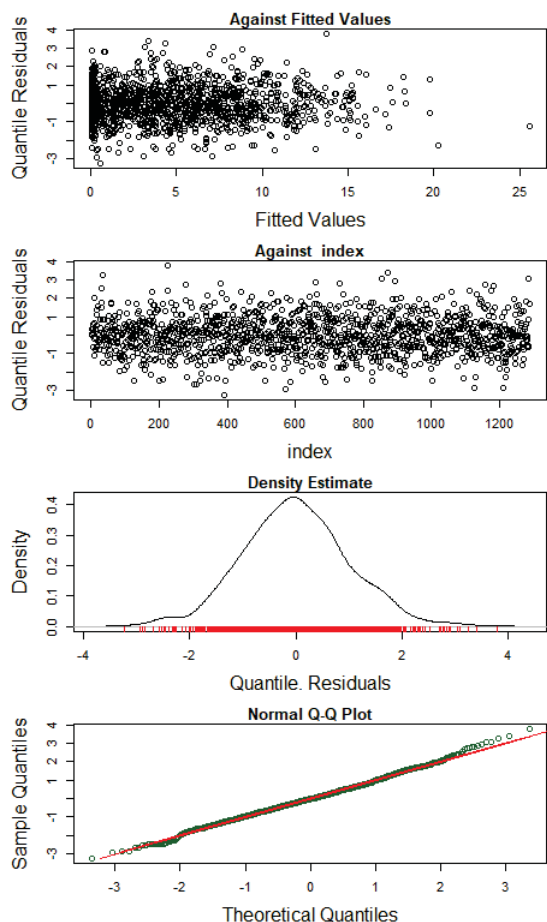


Figura 9: Gráficos da análise dos resíduos.

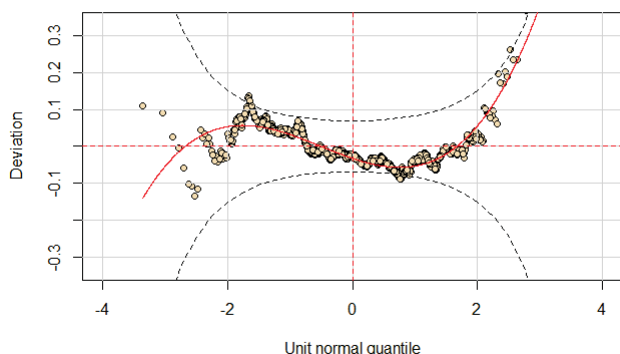


Figura 10: Worm plot para os resíduos do modelo.

5 Conclusão

O objetivo inicial de produzir um modelo que se ajustasse aos dados de telemetria, para se analisar a quantidade de manutenções corretivas através deles foi atingido com êxito. O tipo de regressão escolhida, GAMLSS, se mostrou muito flexível e capaz de se ajustar à com-

plexidade dos dados utilizados. A inclusão de um parâmetro de inflação de zeros, aliado à necessidade de contemplar a superdispersão, foram fundamentais na definição da distribuição a ser utilizada no modelo.

A seleção de covariáveis pelo método *stepwise* está muito bem implementada no pacote GAMLSS e facilita muito na definição dos parâmetros de média, dispersão e inflação de zeros, resultando em um modelo melhor e mais simplificado.

Os suavizadores ajudaram a melhorar o ajuste do modelo construído, ressaltando as relações não lineares entre preditores e resposta através dos gráficos do tipo *term-plot* apresentados nas Figuras 6 a 8.

Os gráficos finais, Figura 9 e 10, mostraram resultados satisfatórios com poucos pontos fora da região de confiança de 95% e fora do comportamento esperado para os resíduos.

O ajuste final do modelo representa uma relação direta de aumento de hodômetro, idade e ociosidade de motor com a frequência de manutenções corretivas. Os gráficos *term-plots* figuras 6 e 8 mostraram uma variância muito grande nos dados dos veículos com hodômetro abaixo de 3000 km (logarítmo natural igual a 8), portanto o modelo é mais preciso para veículos com maior quilometragem.

Uma modificação do modelo poderia ser feita levando em consideração as diversas categorias de veículos além de *hatch* compacto, que foram desconsideradas na seção 2. Com um volume maior de dados dessas outras categorias, poderia ser inserida no modelo uma nova covariável "categoria do veículo". Quanto aos fornecedores, devido à falta de padronização nas configurações dos sensores, ficaria inviável juntá-los em um só modelo, sendo realmente necessária a separação deles.

Referências

- [1] G. Heller R. A. Rigby, M. D. Stasinopoulos and F. De Bastiani. *A Flexible Regression Approach Using GAMLSS in R*. Chapman and Hall/CRC, 2017.
- [2] Teletrac Navman. What is telematics? <https://www.teletracnavman.com/resources/resource-library/faqs/what-is-telematics>, Acesso em: 2020-09-11.
- [3] Telematics.com. What is a fleet tracking system? how does it work? <https://www.telematics.com/a-comprehensive-guide-to-fleet-tracking-systems/#ch1>, Acesso em: 2020-09-11.
- [4] AUTOO. Soluções de mobilidade. categorias. <https://www.autoo.com.br/categorias/>, Acesso em: 2020-05-10.
- [5] G. Heller R. A. Rigby, M. D. Stasinopoulos and F. De Bastiani. *Distributions for Modelling Location, Scale and Shape: Using GAMLSS in R*. Chapman and Hall/CRC, 2017.

- [6] stepgaic: Choose a model by gaic in a stepwise algorithm. <https://www.rdocumentation.org/packages/gamlss/versions/5.1-6/topics/stepGAIC>, Acesso em: 2020-09-07.
- [7] C. Akantziliotou R. A. Rigby, M. D. Stasinopoulos. *Instructions on how to use the gamlss package in R*. Chapman and Hall/CRC, 2008. <http://www.gamlss.com/wp-content/uploads/2013/01/gamlss-manual.pdf>.